**Final Project Report**

**I/ Insurance Data**

The “Exploring Data” shows my work with “Insurance Data”. According to my analysis, there is some useful information from this data:

* The number of car brands accounted for the most on the list is Ford (account for ~9.94 % total), following by Chevrolet (~7.71%) and Toyota (~5.78%).
* Most customers choose to pay the $2,500 Deductible option and $1,000 Deductible is the least selected option based on the bar chart.
* In this list, the proportion of customers who have had a car accident is only about 1/10 of the total number of customers. We have nearly 90% of customers who have not had a car accident (Pie chart).

Actually, I experimented with customers on the list who had a car accident. I want to know that with each deductible option (250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500), what is the average annual premium? Looking at "Exploring\_Data" work, you can easily see the number of customers for each deductible option. To do fairness, I chose the first 48 customers in each option as the sample because there are 48 customers in the "$250" option, and this also the smallest group of people compared to the other. After calculating, we have a table showing all the results. Overall, ignoring a few small deviations, we can conclude that if the customers (had accident) wants to get low annual premium, they must accept a deductible option with a high amount.

In addition, I also experimented to know what percentage of the probability is to predict whether the customer has ever experienced a car accident based on the deductible option that person chose and the annual premium they have to pay. We have a scenario. Let's say I work at an insurance company and have a list of clients about their Deductible and Annual Premium at another company. In order to be able to determine how much money they need to pay for my company annual, I need to consider whether they have been in a car accident. I have applied the Logistic Regression method to create a predictive model. The result showed that the coefficients of both Deductible and Annual Premium are positive. Even though the two values are minuscule they're fine. Moreover, The accuracy of applying the model to training and validation in the predictive variable is nearly 90%. This shows that the model is good, so the outcome has quite a high accuracy. However, the percent probability for the prediction is quite small because maybe the coefficients of Deductible and Annual Premium are quite small. According to the histogram chart, it can be concluded that using the Deductible and Annual Premium in historical records to predict the likelihood that a customer has ever experienced a car accident or not is probably not feasible. In fact, only a small number of customers are predictable, but the probability is still quite low, only in the 25% range.

**II/ IMDB Data**

With this dataset, I want to rely on columns in the table to be able to predict the IMDB score. First, for ease of use, I used linear regression and used the column "num\_voted\_users" to create a predictive model. First of all, I have removed the rows that have a null value in a certain column because this will greatly affect the model. Looking at the scatter plot, it can be seen that the points tend to go up, but there are a lot of alien points scattered in positions which are far way regression line. This shows that the correlation between "imdb\_score" and "num\_voted\_users" does not really have high results. After processing the steps, we will get a model like this: imdb\_score = 6.11 +3.35 \* num\_voted\_users. From this model, I have computed the IMDB score and entered the prediction column. After that, I checked how good was my model and calculated the mean of error. The result is 0.93 which is a pretty big difference between the predicted score and the actual IMDB. This is also proved when the correlation value between the predicted point and the actual point is only 0.48. Due to the R2, “num\_voted\_users” accounts for about 23% of the variation in “imdb\_score”. That means that something else is causing the other 77%. It could be random noise or other variables. Therefore, I changed my mind to use the multiple regression. Still repeating the same steps as in the linear regression section, this time the outcomes have a little bit of improvement. The mean of error has dropped to 0.87, but I rate it still high. In addition, the correlation between the score predicted and actually increased only to 0.56 even though I used all the variables that were possible for prediction. Moreover, R2 only increases slightly to 32%. In conclusion, It can be seen that all the variables that can be used in the IMDB score prediction model in this data are not highly correlated with the variable "imdb\_score". Thus, this multiple regression model is probably not a relatively good one because the mean of error is still quite high, and R2 accounts for just over 30%, and the remaining 70% comes from random noise variables.